LETTER CHANGE BIAS AND LOCAL UNIQUENESS IN OPTIMAL SEQUENCE ALIGNMENTS

RAPHAEL HAUSER† AND HEINRICH MATZINGER ‡

Abstract. Considering two optimally aligned random sequences, we investigate the effect on the alignment score caused by changing a random letter in one of the two sequences. Using this idea in conjunction with large deviations theory, we show that in alignments with a low proportion of gaps the optimal alignment is locally unique in most places with high probability. This has implications in the design of recently pioneered alignment methods that use the local uniqueness as a homology indicator.

AMS subject classifications. Primary 60F10, Secondary 92D20.

Key words. Random sequences, sequence alignment, large deviations.

1. Introduction.

The purpose of this paper is to gain insight into the local multiplicity of optimal alignments of two random binary sequences. Before introducing the problem setting, let us give a brief motivation and literature review.

1.1. Motivation.

A fairly general and useful technique to identify high quality alignments of two sequences $x =' x_1 \dots x_m'$ and $y =' y_1 \dots y_n'$ with characters from a finite alphabet $\mathbb{A}$ is to consider alignments with gaps $\sqcup$,

$$\begin{array}{ccccc} \sqcup & x_1 & x_2 & x_3 & \sqcup \\ y_1 & y_2 & \sqcup & y_3 & y_4 \end{array}$$

and to quantify the quality of such an alignment with a score of the form

$$S(x, y) = s(\sqcup, y_1) + s(x_1, y_2) + s(x_2, \sqcup) + s(x_3, y_3) + s(\sqcup, y_4). \tag{1.1}$$

When a scoring function of the form (1.1) is used, a good choice of individual scores $s(a, b)$ of matched symbols a and b and the choice of the alphabet $\mathbb{A}$ depend on the specific application one has in mind. The matching of any symbol $a \in \mathbb{A}$ with a gap $\sqcup$ is typically penalized by a negative score term $s(a, \sqcup)$, $s(\sqcup, a) < 0$.

The simplest similarity measure of the form (1.1) arises as the length of a *longest common subsequence* (LCS). A common subsequence is any sequence that can be obtained by deleting some characters of either sequence and keeping the remaining ones in the original order. The length of a longest common subsequence of two sequences x and y is the same as the score $S(x, y)$, where individual scoring terms are defined as follows,

$$s(a, b) = \begin{cases} 1 & \text{if } a = b \neq \sqcup, \\ \infty & \text{if } a \neq b \text{ and } a, b \neq \sqcup, \\ 0 & \text{if } a = \sqcup \text{ or } b = \sqcup, \text{ but not both.} \end{cases}$$

Sequence alignment techniques play an important role in biology (see e.g. [21]), speech recognition, pattern recognition, automated translation and other areas where

†Mathematical Institute, 2429 St Giles', Oxford OX1 3LB, United Kingdom. (hauser@maths.ox.ac.uk). This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/H02686X/1].

‡School of Mathematics, Georgia Institute of Technology, 686 Cherry Street, Atlanta, GA 30332-0160, USA, (matzi@math.gatech.edu).

hidden Markov models are used as an analytic tool. The study of optimal alignments of random sequences is also interesting to statistical physicists, because it can be seen as a first passage percolation problem on an oriented graph with correlated weights. First passage percolation is a mature field and has been a major research area in discrete probability for several decades. An excellent overview can be found in the chapter dedicated to first passage percolation of Volume 110 of the Springer Encyclopaedia of Mathematical Sciences.

Among several long-standing open questions in this field are the problems of identifying the exact order of the fluctuations and the proportion of points where the optimal alignment is unique. Recently, significant progress has been made on both questions: In several special cases it was shown that a positive bias effect of a random letter-change on the optimal alignment score of two random sequences of length n exists and implies an order of fluctuation proportional to $\sqrt{n}$, see [8], [17], [15], [16]. In [3] it was shown how to apply this principle to arbitrary scoring functions, and how to detect the positive bias effect via a nontrivial Montecarlo technique. In [13], a case study was conducted on using the local uniqueness of optimal alignments to detect the homology of two DNA sequences. The motivation behind this method is the empirical observation that all optimal alignments of two biologically related sequences are identical in most places, while optimal alignments of unrelated sequences are locally nonunique in most places.

Our paper concerns a theoretical study of this last observation. Considering two independent random sequences with i.i.d. characters, we examine their optimal alignments containing a fixed proportion of gaps and prove that when the proportion of gaps is small, then with high probability optimal alignments differ only in a small number of places and are locally unique everywhere else. Our result implies that the approach of [13] can only work for scoring functions in which gaps are not penalized too heavily, as this would force the number of gaps appearing in optimal alignments to be small relative to the length of the two sequences.

Optimal alignments of random sequences are often used as null-models in statistical tests to decide on whether two or more given sequences are homologous. The mathematical underpinnings are best understood in the context of the LCS setting. Let L_n denote the length of the LCS of two independent binary i.i.d. sequences of length n. Using a subadditivity argument, Chvàtal-Sankoff [9] showed that the limit

$$\gamma := \lim_{n\to\infty} \frac{E[L_n]}{n}$$

exists. Determining the exact value of γ – the so-called *Chvàtal-Sankoff constant* – is a long standing open problem. However, upper and lower bounds are known, see [9], [6], [11], [10, 18], [12], [2].

Another long open problem is the determination of the exact order of the fluctuation of the LCS score as a function of the length of the sequences. Considering the case of binary sequences obtained by flipping perfect coins, it was shown in [20] that $VAR[L_n] \leq n$. Montecarlo simulations in [9] led to the conjecture that $VAR[L_n] = o(n^{2/3})$. This order of magnitude is similar to the order for the so-called *longest increasing subsequence* (LIS) of random permutations, see [7] and [1]). The LIS setting is asymptotically equivalent to first passage percolation on a oriented Poisson random graph. In [22] it was conjectured that in many cases the variance of L_n grows linearly. This seems indeed to be the case generically [3], but there may exist different orders of magnitude for these fluctuations, depending on the distribution of the sequences X and Y, see [8], [17], [4].

1.2. Problem Setting and Key Ideas. Let us consider the set of optimal LCS alignments of the two sequences 'I-do-not-like-symmetry' and 'I-detest-symmetry'. If we were to give an exhaustive list of optimal alignments, we would find that all of them are of the form

$$\begin{array}{cccccccccccccc} I & - & d & * & \dots & * & - & s & y & m & m & e & t & r & y \\ I & - & d & * & \dots & * & - & s & y & m & m & e & t & r & y, \end{array}$$

that is, all optimal LCS alignments agree in the region outside of the places filled with wild card stars $*$. We say that in this region the optimal alignment is *locally unique*. In our example the optimal LCS alignment is locally unique on a large proportion of the two sequences.

This observation is typical not only in the LCS setting, but for general scoring functions $S(x, y)$ as introduced in (1.1): When two sequences are closely related to one another, the optimal alignments are locally unique in many places. Conversely, the optimal alignments of two random sequences with i.i.d. entries often are locally unique only in very few places. [13] exploited this observation in the design of a homology detecting algorithm for DNA sequences. However, as this paper will reveal, for this method to work one has to select the gap penalty $s(a, \sqcup)$ quite carefully: When gaps are strongly penalized, then no more than a constant proportion of gaps are observed in optimal alignments, and if the proportion of gaps is small then the optimal alignment is locally unique in most places even for i.i.d. random (and hence totally unrelated) sequences.

Our paper concerns a theoretical analysis of this phenomenon. To make the analysis transparent, we chose a simplified setup in which $X = X_1 \dots X_m$ and $Y = Y_1 \dots Y_n$ are random sequences consisting of i.i.d. standard Bernoulli variables, where $m = \lfloor (1-\delta)n \rfloor$ depends on n via a fixed gap proportion δ. We then investigate alignments of X and Y that contain gaps only in X, and we use a scoring function in which matching symbols contribute to the total score with unit weight and non-matching ones with zero weight. Our interest is in the random number $U \leq m$ of indices i for which X_i is aligned with more than one Y_j under the different optimal alignments. The main theorem of this paper shows that when δ is small and n is large, optimal alignments are locally unique in an arbitrarily large proportion of places with arbitrarily high probability:

THEOREM 1.1. *For all $\varepsilon > 0$ there exists $\delta_0 > 0$ and $n_0 \in \mathbb{N}$ such that for all $\delta \in (0, \delta_0)$ and $n > n_0$, $\mathrm{P}\left[U \geq m\varepsilon\right] < \varepsilon$.*

While it is clear that $\mathrm{P}[U > m\varepsilon] = 0$ when $\delta = 0$ for any n, the theorem shows the nontrivial fact that the limit $\lim_{n\to\infty} \mathrm{P}[U > m\varepsilon]$ is continuous in δ at $\delta = 0$. Furthermore, the proof provides the quantitative estimate

$$\mathrm{P}\left[U \geq m\varepsilon\right] \leq \frac{\mathcal{O}\left(\delta^{\frac{1}{2}}\right)}{\mathcal{O}(\varepsilon) + \mathcal{O}\left(\delta^{\frac{1}{2}}\right)} + \mathcal{O}\left(\mathrm{e}^{-n\delta}\right).$$

Theorem 1.1 is also very interesting in the context of the Chvàtal-Sankoff conjecture which concerns the order of magnitude of the fluctuation of the LCS of two random texts.

We recall the assumptions under which we prove our main result and which will remain valid throughout the rest of this paper: Let $n \in \mathbb{N}$ and let $0 < \delta < 1$ be a fixed constant not depending on n. We set $m = \lfloor n - \delta n \rfloor$ and define two independent random

sequences $X = X_1 \dots X_m$ and $Y = Y_1 \dots Y_n$ by choosing $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ as i.i.d. Bernoulli variables with parameter 1/2 (i.e., coin tossing experiments). We then consider alignments

$$\begin{array}{ccccccccccc} \sqcup & \dots & \sqcup & X_1 & \sqcup & \dots & \sqcup & X_m & \sqcup & \dots & \sqcup \\ Y_1 & \dots & Y_{\xi(1)-1} & Y_{\xi(1)} & Y_{\xi(1)+1} & \dots & Y_{\xi(m)-1} & Y_{\xi(m)} & Y_{\xi(m)+1} & \dots & Y_n \end{array}$$

with gaps in X only and attribute to it the score

$$S(X, Y; \xi) = \#\{i \in \mathbb{N}_m : X_i = Y_{\xi(i)}\},$$

where # denotes the cardinality of a set and $\mathbb{N}_n := \{1, \dots, n\}$. The set of alignments ξ that maximize $S(X, Y; \xi)$ is denoted by $\mathcal{O}A_{X,Y}$. Of course, this is a random set of alignments, since X and Y are random. We write $S^*(x, y) := \max_\xi S(X, Y; \xi)$ for the maximum score and

$$U := \#\{i \in \mathbb{N}_m : \exists \xi, \lambda \in \mathcal{O}A_{X,Y} \text{ s.t. } \xi(i) \neq \lambda(i)\}$$

for the number of positions where X is not uniquely aligned with Y among the alignments with maximum score. U is a random variable.

Theorem 1.1 states that $\mathrm{P}[U \geq m\varepsilon] < \varepsilon$ for all n large enough and δ small enough. In other words, for large n and small gap proportion δ the optimal alignment is locally unique in a $(1 - \epsilon)$-proportion of the sequence X with probability greater than $1 - \epsilon$. This result is qualitatively representative for what occurs with regards to the local uniqueness of optimal alignments of random sequences under *arbitrary* scoring functions $S(X, Y)$ as defined in (1.1) whenever gaps are strongly penalized, i.e., $s(a, \sqcup)$ is a negative number of not too small a modulus. Indeed, strong gap penalization prevents optimal alignments from having more than a small proportion of gaps. Furthermore, allowing gaps only in one of the sequences is merely a technical assumption that vastly simplifies the analysis.

Let us now briefly explain the main idea behind the proof of Theorem 1.1. We define a measure preserving map by picking an entry of X at random and flipping it to the "opposite" value, i.e., a 0 is changed to a 1 or vice versa (we imagine X as a line-up of randomly tossed fair coins). We denote the sequence obtained in this fashion by $\tilde{X}$. Since this operation is measure preserving, we have

$$\mathrm{E}[\Delta] = 0, \tag{1.2}$$

where $\Delta := S^*(\tilde{X}, Y) - S^*(X, Y)$. A crucial observation is now that when the optimal alignment is nonunique in a large proportion of places, then the optimal score tends to increase under this measure-preserving map. We illustrate this phenomenon in Example 1.2 below. Together with 1.2 this observation implies that the probability that the optimal alignment is nonunique in many points is small.

Example 1.2. *Consider the case where $n = 8$, $m = 6$, $\delta = 1/4$ and X and Y take the values $x = 001110$ and $y = 11110011$. There are two optimal alignments, ξ given by $\xi(i) = i$ for $(i = 1, \dots, 6)$ or*

$$\begin{array}{cccccccc} 0 & 0 & 1 & 1 & 1 & 0 & \sqcup & \sqcup \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1, \end{array}$$

and λ given by $\lambda(i) = i$ for $(i = 1, \dots, 4)$ and $\lambda(5) = 7$, $\lambda(6) = 8$ or

$$\begin{array}{cccccccc} 0 & 0 & 1 & 1 & \sqcup & \sqcup & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1. \end{array}$$

The optimal score is $S^*(x, y) = S(x, y; \xi) = S(x, y; \lambda) = 3$.

The following combinatorial property holds for arbitrary alignments ξ, λ of x and y, not only optimal ones: If $i \in \{1, \dots, m\}$ is such that

$$y_{\xi(i)} \neq y_{\lambda(i)} \tag{1.3}$$

then flipping x_i to the opposite value increases at least one of the scores $S(x, y; \xi)$, $S(x, y; \lambda)$ by one unit. In particular, if ξ and λ are both optimal alignments and condition (1.3) holds, then flipping x_i to the opposite value increases the optimal score by one unit. For the chosen values of x and y we find that $i = 5, 6$ satisfy condition (1.3). Flipping the value of x_5 from 1 to 0, we find that the score of ξ increases to 4 and the score of λ decreases to 2. The maximum score is now 4. Similarly, flipping x_6 from 0 to 1, the score of ξ decreases to 2 whereas the score of λ increases to 4. The new maximum score is again 4. If an entry x_T of x is flipped, where T is a random index in $\mathbb{N}_m$, then this implies

$$\mathrm{E}[\Delta \| X = x, Y = y, \xi(T) \neq \lambda(T), Y_{\xi(T)} \neq Y_{\lambda(T)}] = 1. \tag{1.4}$$

On the other hand, if one of the entries $x_1, \dots, x_4$ that do not satisfy (1.3) is flipped, then the maximum score can either increase or decrease: For $i = 1 \dots, 4$, we have $\xi(i) = \lambda(i)$. The entries x_1 and x_2 are aligned with non-matching symbols, so that flipping one of these entries increases the optimal score by one. The entries x_3 and x_4 are aligned with matching symbols. In the present example, switching one of these entries results in a decrease of the optimal score by one unit, though in other cases the maximum score can remain unchanged (but it will then be attained by a different alignment). Thus, we find that if a random entry x_T of x is flipped (where $T \in \mathbb{N}_m$) then for the above choices of x and y,

$$\mathrm{E}[\Delta \| X = x, Y = y, \xi(T) = \lambda(T)] \geq 0. \tag{1.5}$$

For the same reason, if there were any indices i such that $\xi(i) \neq \lambda(i)$ where (1.3) does not hold, then we would find

$$E[\Delta \| X = x, Y = y, \xi(T) \neq \lambda(T), Y_{\xi(T)} = Y_{\lambda(T)}] \geq 0.$$

In conjuction with (1.4) this implies

$$\begin{aligned} &\mathrm{E}[\Delta \| X = x, Y = y, \xi(T) \neq \lambda(T)] \\ &\qquad \geq \mathrm{P}[Y_{\xi(T)} \neq Y_{\lambda(T)} \| X = x, Y = y, \xi(T) \neq \lambda(T)]. \end{aligned} \tag{1.6}$$

The proof of Theorem 1.1 exploits a generalization of these observations: Lemma 2.3 of Section 2 shows that there exist two optimal alignments ξ and λ that differ from each other exactly in positions i where the optimal alignment of X and Y is not locally unique. In Section 3 we show that up to negatively exponentially small probability in n the following are true,

i) approximately half the points $i \in \mathbb{N}_m$ with $\xi(i) = \lambda(i)$ satisfy $X_i \neq Y_{\lambda(i)}$,
ii) approximately half the points $i \in \mathbb{N}_m$ with $\xi(i) \neq \lambda(i)$ satisfy $Y_{\xi(i)} \neq Y_{\lambda(i)}$,
iii) approximately a quarter of points $i \in \mathbb{N}_m$ with $\xi(i) \neq \lambda(i)$ satisfy $X_i \neq Y_{\xi(i)} = Y_{\lambda(i)}$,
iv) approximately a quarter of points $i \in \mathbb{N}_m$ with $\xi(i) \neq \lambda(i)$ satisfy $X_i = Y_{\xi(i)} = Y_{\lambda(i)}$,

v) for all $e_1, e_2 \in \{0,1\}$ approximately a quarter of points $i \in \mathbb{N}_m$ satisfy $X_i = e_1$ and $Y_{\xi(i)} = e_2$.

Let T be the random index in $\mathbb{N}_m$ that corresponds to the entry of X that is flipped. By the observations of Example 1.2, i)–iv) lead to the following generalization of (1.6),

$$\mathrm{E}[\Delta \| X = x, Y = y, \xi(T) \neq \lambda(T)] \geq \frac{1}{2}. \tag{1.7}$$

Likewise, v) leads to the following generalization of (1.5),

$$\mathrm{E}[\Delta \| X = x, Y = y, \xi(T) = \lambda(T)] \geq 0. \tag{1.8}$$

Of course, (1.7) and (1.8) hold only approximately. Much of the work of Section 3 is devoted to overcoming these imprecisions. For now, let us work with the simplified assumption that (1.7) and (1.8) hold true except on a set F^c of pairs (X, Y) with negatively exponentially small probability $\mathrm{P}[F^c] = \exp(-\mathcal{O}(n))$. Equations (1.2), (1.7) and (1.8) then imply

$$0 = \mathrm{E}[\Delta] \geq \frac{1}{2} \times \mathrm{P}[\xi(T) \neq \lambda(T)] + 0 \times \mathrm{P}[\xi(T) = \lambda(T)] - 1 \times \mathrm{P}[F^c],$$

so that

$$\mathrm{P}[\xi(T) \neq \lambda(T)] \leq \exp(-\mathcal{O}(n)). \tag{1.9}$$

When the approximate statements (1.7) and (1.8) are replaced with correct inequalities, (1.9) turns into the weaker claim of Theorem 1.1.

The structure of the remaining sections of this paper is as follows. Section 2 serves to introduce the main notation relevant to scoring functions, alignments and local uniqueness of alignments. We also discuss illustrative examples and prove two technical results of preliminary nature. In Section 3 we introduce events defined in terms of certain empirical distributions and their large deviations. These events allow putting the above-made approximate statements i)–v) onto a rigorous footing. In Section 4 we define formally the measure-preserving map which flips a random entry of X to its opposite value. In Lemma 4.1 of that section we prove that the locations where the optimal alignment is nonunique tend to introduce a positive bias into $\mathrm{E}[\Delta]$. Section 5 finally brings all the elements together in the proof of Theorem 1.1.

2. Alignments and Scores. Let $(x, y) \in \{0,1\}^m \times \{0,1\}^n$ be a pair of sequences of lengths $m < n$ over the binary alphabet. Let us consider alignments

$$\begin{array}{ccccccccccc} y_1 & \cdots & y_{\xi(1)-1} & y_{\xi(1)} & y_{\xi(1)+1} & \cdots & y_{\xi(m)-1} & y_{\xi(m)} & y_{\xi(m)+1} & \cdots & y_n \\ \sqcup & \ldots & \sqcup & x_1 & \sqcup & \ldots & \sqcup & x_m & \sqcup & \ldots & \sqcup \end{array}$$

of x and y that have δn gaps in x, where $m = \lfloor (1-\delta)n \rfloor$. In the above display we marked gaps with the symbol $\sqcup$. We identify the set of such alignments with the set $\mathcal{A}_{m,n}$ of order-preserving injections of $\mathbb{N}_m$ into $\mathbb{N}_n := \{1, \ldots, n\}$, that is, $\xi \in \mathcal{A}_{m,n}$ if and only if $\xi : \mathbb{N}_m \hookrightarrow \mathbb{N}_n$ and $i < j$ implies $\xi(i) < \xi(j)$.

Lemma 2.1. *If $\delta < 5/6$, then $\#\mathcal{A}_{m,n} \leq \exp\big(nH(\delta)\big)$, where $H(\delta) = -\big(\delta \ln \delta + (1-\delta)\ln(1-\delta)\big)$ is the entropy function.*

Proof. Robbins' inequality [19] says that

$$\sqrt{2\pi} n^{-n+\frac{1}{12+1}} \leq n! \leq \sqrt{2\pi} n^{n+\frac{1}{2}} \, \mathrm{e}^{-n+\frac{1}{12n}} .$$

Therefore,

$$\begin{aligned}
\#\mathcal{A}_{m,n} &= \binom{n}{n(1-\delta)} \\
&\leq \frac{\sqrt{2\pi} n^{n+\frac{1}{2}} \, \mathrm{e}^{-n+\frac{1}{12n}}}{\sqrt{2\pi}\big(n(1-\delta)\big)^{n(1-\delta)+\frac{1}{2}} \, \mathrm{e}^{-n(1-\delta)+\frac{1}{12n(1-\delta)+1}} \times \sqrt{2\pi}(n\delta)^{n\delta+\frac{1}{2}} \, \mathrm{e}^{-n\delta+\frac{1}{12n\delta+1}}} \\
&\qquad\qquad = \mathrm{e}^{nH(\delta)} \times \frac{\exp\left(\frac{1}{12n} - \frac{1}{12n(1-\delta)+1} - \frac{1}{12n\delta+1}\right)}{\sqrt{2\pi(1-\delta)(n-m)}}.
\end{aligned}$$

Note that the second factor converges to zero for fixed δ and $n \to \infty$. Moreover, the numerator is < 1 and for $\delta < 5/6$ the denominator is > 1 since $2\pi(1-\delta)(n-m) > 6(1-\delta) > 1$. □

We define a scoring function $\{0,1\}^m \times \{0,1\}^n \times \mathcal{A}_{m,n} \to \mathbb{N}_0$ as follows,

$$S(x, y; \xi) := \sum_{i=1}^{m} s(x_i, y_{\xi(i)}),$$

where $s(0,0) = s(1,1) = 1$ and $s(0,1) = s(1,0) = 0$. The set of optimal alignments of (x, y) is the set of alignments with maximum score,

$$\mathcal{O}A_{x,y} := \{\xi \in \mathcal{A}_{m,n} : S(x,y;\xi) \geq S(x,y;\lambda) \ \forall \lambda \in \mathcal{A}_{m,n}\}.$$

We write $S^*(x,y) := \max\{S(x,y;\xi) : \xi \in \mathcal{A}_{m,n}\}$ for the maximum score, as before.

For each $i \in \mathbb{N}_m$ we define the variable

$$u_i(x,y) := \begin{cases} 1 & \text{if } \exists \xi, \lambda \in \mathcal{O}A_{x,y} \text{ s.t. } \xi(i) \neq \lambda(i), \\ 0 & \text{otherwise} \end{cases}$$

that indicates when the image of i under the set of optimal alignments is nonunique. We say that the optimal alignment is *locally nonunique* at i if $u_i(x,y) = 1$. We write

$$u(x,y) := \sum_{i=1}^{m} u_i(x,y)$$

for the number of indices where the optimal alignment is locally nonunique.

The sets $\mathcal{O}A_{x,y}$ and $\{i \in \mathbb{N}_m : u_i(x,y) = 1\}$ can be found via dynamic programming: A $m \times n$ matrix $\big(\mathit{score}(i,j)\big)$ is recursively computed, using the rules

r.i) $\mathit{score}(i,j) = -1$ for $i > j$ or $j > i + \delta n$,

r.ii) $\mathit{score}(1,j) = s(x_1, y_j)$ for $j = 1, \dots, 1 + \delta n$,

r.iii) $\mathit{score}(i,j) = s(x_i, y_j) + \max\{\mathit{score}(i-1,k) : k < j\}$ for all other (i,j).

Arguing recursively, one immediately verifies that

$$S^*(x,y) = \max\{\mathit{score}(m,j) : j \in \mathbb{N}_n\},$$

and furthermore that $\xi \in \mathcal{O}A_{x,y}$ if and only if the following conditions are satisfied:

c.i) $\xi(m) \in \big\{j \in \mathbb{N}_n : \mathit{score}(m,j) = S^*(x,y)\big\}$,

c.ii) $\xi(i-1) \in \big\{j < \xi(i) : \mathit{score}(i-1,j) = \max_{k<\xi(i)} \mathit{score}(i-1,k)\big\}$ for all $i = 2, \dots, m$.

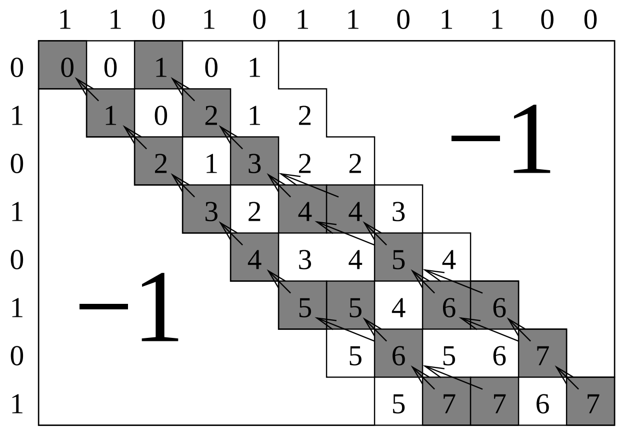

FIG. 2.1. *The scoring matrix of Example 2.2.*

EXAMPLE 2.2. *Let $x = 01010101$ and $y = 110101101100$. Then the above described dynamic programming algorithm generates the matrix of Figure 2.1, where it is displayed in tableau format. Optimal paths follow the arrows and pass through the shaded entries. The tableau is annotated with the generating sequences x and y, so that the optimal alignments can easily be read off. The following table lists y in the top row, followed by a complete list of optimal alignments of x with y,*

1	**1**	**0**	**1**	**0**	**1**	**1**	**0**	**1**	**1**	**0**	**0**
0	1	0	1	0	1	$\sqcup$	0	1	$\sqcup$	$\sqcup$	$\sqcup$
0	1	0	1	0	1	$\sqcup$	0	$\sqcup$	1	$\sqcup$	$\sqcup$
0	1	0	1	0	$\sqcup$	1	0	1	$\sqcup$	$\sqcup$	$\sqcup$
0	1	0	1	0	$\sqcup$	1	0	$\sqcup$	1	$\sqcup$	$\sqcup$
$\sqcup$	$\sqcup$	0	1	0	1	$\sqcup$	0	1	$\sqcup$	0	1
$\sqcup$	$\sqcup$	0	1	0	1	$\sqcup$	0	$\sqcup$	1	0	1
$\sqcup$	$\sqcup$	0	1	0	$\sqcup$	1	0	1	$\sqcup$	0	1
$\sqcup$	$\sqcup$	0	1	0	$\sqcup$	1	0	$\sqcup$	1	0	1

Every line of the the tableau in Figure 2.1 contains multiple shaded entries. Therefore, $u_i(x, y) = 1$ for all $i \in \mathbb{N}_m$ and $u(x, y) = m$.

In the above example we ordered the optimal alignments from leftmost to rightmost as located within the tableau, that is, alignments are listed in inverse alphabetical order with respect to the lateness of gaps. This also provides the idea of proof for the following result, which shows that there exist two optimal alignments that differ from one another at every point where the optimal alignment is locally nonunique.

LEMMA 2.3. *For all $(x, y) \in \{0, 1\}^m \times \{0, 1\}^n$ there exist $\xi, \lambda \in \mathcal{O}A_{x,y}$ such that $\xi(i) \neq \lambda(i)$ for exactly those indices $i \in \mathbb{N}_m$ for which $u_i(x, y) = 1$.*

Proof. The claim is clearly true if we can prove that $\xi, \lambda \in \mathcal{O}A_{x,y}$, where

$$\xi(i) := \min\{\psi(i) : \psi \in \mathcal{O}A_{x,y}\} \quad \forall i \in \mathbb{N}_m,$$
$$\lambda(i) := \max\{\psi(i) : \psi \in \mathcal{O}A_{x,y}\} \quad \forall i \in \mathbb{N}_m.$$

If $\lambda \notin \mathcal{O}A_{x,y}$ then there exists an index $i \in \mathbb{N}_m \setminus \{1\}$ such that $\widehat{\psi}(i-1) < \lambda(i-1)$ for all $\widehat{\psi} \in \mathcal{O}A_{x,y}$ such that $\widehat{\psi}(i) = \lambda(i)$. On the other hand, there exists $\widehat{\lambda} \in \mathcal{O}A_{x,y}$ such

that $\widehat{\lambda}(i-1) = \lambda(i-1)$, and therefore it is necessarily the case that $\widehat{\lambda}(i) < \widehat{\psi}(i) = \lambda(i)$. But $\widehat{\lambda}$ and $\widehat{\psi}$ satisfy condition c.ii), that is,

$$\widehat{\lambda}(i-1) \in \Big\{ j < \widehat{\lambda}(i) : \, score(i-1, j) = \max_{k<\widehat{\lambda}(i)} score(i-1, k) \Big\},$$
$$\widehat{\psi}(i-1) \in \Big\{ j < \widehat{\psi}(i) : \, score(i-1, j) = \max_{k<\widehat{\psi}(i)} score(i-1, k) \Big\}.$$

Therefore, either $\max_{k<\widehat{\lambda}(i)} score(i-1, k) = \max_{k<\widehat{\psi}(i)} score(i-1, k)$ and then there exists $\psi \in \mathcal{O}A_{x,y}$ such that $\psi(i) = \widehat{\psi}(i) = \lambda(i)$ and $\psi(i-1) = \widehat{\lambda}(i-1) = \lambda(i-1)$, or else $\max_{k<\widehat{\lambda}(i)} score(i-1, k) < \max_{k<\widehat{\psi}(i)} score(i-1, k)$ and then $\widehat{\psi}(i-1) > \widehat{\lambda}(i-1) = \lambda(i-1)$. In either case we have a contradiction, and this shows that $\lambda \in \mathcal{O}A_{x,y}$ indeed. The proof that $\xi \in \mathcal{O}A_{x,y}$ is analogous. $\square$

Note that the alignments ξ and λ constructed in the proof of Lemma 2.3 are uniquely determined by (x, y). Furthermore, they satisfy the relation $\xi \leq \lambda$ which is defined by $\xi(i) \leq \lambda(i)$ for all $i \in \mathbb{N}_m$.

3. Large Deviations of Some Empirical Distributions. In this section we establish a rigorous version of the approximate inequalities (1.7),(1.8) and statements i)–v) of Section 1.2. Recall that $X = X_1 \dots X_m$ and $Y = Y_1 \dots Y_n$ are two independent random sequences that consist of i.i.d. standard Bernoulli variables $X_i, Y_j \sim \mathscr{B}(1/2)$ defined on some probability space $(\Omega, \mathscr{F}, \mathrm{P})$. We write U_i and U for the random variables $u_i(X, Y)$ and $u(X, Y)$ respectively. Furthermore, we think of $\delta \in (0, 1)$ as a fixed gap proportion that relates m to n via $m = \lfloor (1-\delta)n \rfloor$.

For $\xi \in \mathcal{A}_{m,n}$ fixed and $\omega \in \Omega$ let $\widehat{\mathscr{D}}_\xi(\omega)$ be the empirical distribution of $(X_i(\omega), Y_{\xi(i)}(\omega))$ over $i \in \mathbb{N}_m$, i.e., the distribution of $(X_T(\omega), Y_{\xi(T)}(\omega))$ when $T \sim \mathscr{U}(\mathbb{N}_m)$ is a random index with uniform distribution on $\mathbb{N}_m$. Yet another way to define this distribution is to require that for all $(e_1, e_2) \in \{0, 1\}^2$,

$$\mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(e_1, e_2)\big] = \frac{1}{m} \times \#\Big\{ i \in \mathbb{N}_m : \, X_i(\omega) = e_1, \, Y_{\xi(i)}(\omega) = e_2 \Big\}.$$

Example 3.1. *Let x, y and ξ be chosen as in Example 1.2. If $\omega \in \Omega$ is chosen such that $X(\omega) = x$ and $Y(\omega) = y$ then*

$$\mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(0,0)\big] = \frac{1}{6}, \; \mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(0,1)\big] = \frac{2}{6}, \; \mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(1,0)\big] = \frac{1}{6}, \; \mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(1,1)\big] = \frac{2}{6}.$$

Note that $\widehat{\mathscr{D}}_\xi$ only depends on x and y.

Let E_ξ be the event that

$$\max_{(e_1,e_2)\in\{0,1\}^2} \Big| \mathrm{P}_{\widehat{\mathscr{D}}_\xi(\omega)}\big[(e_1, e_2)\big] - 1/4 \Big| < \sqrt{\frac{9H(\delta)}{4(1-\delta)}} =: \epsilon_1(\delta), \tag{3.1}$$

in other words,

$$E_\xi := \Big\{ \omega \in \Omega : \, \Big\| \widehat{\mathscr{D}}_\xi(\omega) - \mathscr{B}(1/2) \otimes \mathscr{B}(1/2) \Big\| < \epsilon_1(\delta) \Big\}.$$

Let us furthermore define the event

$$E_{m,n} := \bigcap_{\xi \in \mathcal{A}_{m,n}} E_\xi.$$

In a similar vein we define empirical distributions and events relating to $(X_i, Y_{\xi(i)}, Y_{\lambda(i)})$ as follows: Let $\varepsilon > 0$ and $\xi, \lambda \in \mathcal{A}_{m,n}$ be fixed such that $\xi \leq \lambda$ and $d(\xi, \lambda) := \#\{i : \xi(i) \neq \lambda(i)\} \geq m\varepsilon$. For all i such that $\xi(i) = \lambda(i)$ we define the random variables

$$R_i^{\xi,\lambda} := \begin{cases} 1 & \text{if } X_i \neq Y_{\xi(i)} \\ -1 & \text{if } X_i = Y_{\xi(i)}. \end{cases}$$

Likewise, for all i such that $\xi(i) \neq \lambda(i)$ we define the random variables

$$R_i^{\xi,\lambda} := \begin{cases} 0 & \text{if } Y_{\xi(i)} \neq Y_{\lambda(i)}, \\ 1 & \text{if } X_i \neq Y_{\xi(i)} = Y_{\lambda(i)}, \\ -1 & \text{if } X_i = Y_{\xi(i)} = Y_{\lambda(i)}. \end{cases}$$

We now introduce the following notation:

- $\widehat{\mathscr{L}}_{\xi,\lambda}^{agree}(\omega)$, $\widehat{\mathscr{L}}_{\xi,\lambda}^{disag}(\omega)$ and $\widehat{\mathscr{L}}_{\xi,\lambda}^{unif}(\omega)$ are the empirical distributions of $R_i^{\xi,\lambda}(\omega)$ over $\{i \in \mathbb{N}_m : \xi(i) = \lambda(i)\}$, $\{i \in \mathbb{N}_m : \xi(i) \neq \lambda(i)\}$ and $i \in \mathbb{N}_m$ respectively,
- $\mathscr{J}^{agree}$ is the distribution on $\{-1, 1\}$ defined by $\mathrm{P}_{\mathscr{J}^{agree}}[-1] = 1/2 = \mathrm{P}_{\mathscr{J}^{agree}}[1]$, and $\mathscr{J}^{disag} = \mathscr{J}^{unif}$ the distribution on $\{-1, 0, 1\}$ defined by $\mathrm{P}_{\mathscr{J}^{disag}}[-1] = 1/4 = \mathrm{P}_{\mathscr{J}^{disag}}[1]$ and $\mathrm{P}_{\mathscr{J}^{disag}}[0] = 1/2$,
- parameters ϵ_2, ϵ_3 and ϵ_4 are determined in terms of δ and ε via the relations

$$\epsilon_2(\delta, \varepsilon) := \sqrt{\frac{3H(\delta)}{2(1-\delta)\varepsilon}}, \quad \epsilon_3(\delta, \varepsilon) := \sqrt{\frac{27H(\delta)}{8(1-\delta)\varepsilon}}, \quad \epsilon_4(\delta, \varepsilon) := \sqrt{\frac{3H(\delta)}{2(1-\delta)(1-\varepsilon)}}.$$

With this notation, we define the following events,

$$\begin{aligned} F_{\xi,\lambda}^{\varepsilon} &:= \left\{\omega \in \Omega : \left\|\widehat{\mathscr{L}}_{\xi,\lambda}^{agree} - \mathscr{J}^{agree}\right\| < \epsilon_2(\delta, \varepsilon)\right\}, \\ G_{\xi,\lambda}^{\varepsilon} &:= \left\{\omega \in \Omega : \left\|\widehat{\mathscr{L}}_{\xi,\lambda}^{disag} - \mathscr{J}^{disag}\right\| < \epsilon_3(\delta, \varepsilon)\right\}, \\ H_{\xi,\lambda}^{\varepsilon} &:= \left\{\omega \in \Omega : \left\|\widehat{\mathscr{L}}_{\xi,\lambda}^{unif} - \mathscr{J}^{unif}\right\| < \epsilon_4(\delta, \varepsilon) + 2\varepsilon\right\}, \\ F_{m,n,\varepsilon} &:= \bigcap_{\{(\xi,\lambda):\xi\leq\lambda, m\varepsilon\leq d(\xi,\lambda)\leq m(1-\varepsilon)\}} \left(F_{\xi,\lambda}^{\varepsilon} \cap G_{\xi,\lambda}^{\varepsilon}\right) \quad \cap \bigcap_{\{(\xi,\lambda):\xi\leq\lambda, m(1-\varepsilon)<d(\xi,\lambda)\}} H_{\xi,\lambda}^{\varepsilon}. \end{aligned}$$

Lemma 3.2. *For all $\delta < 5/6$ and $\varepsilon > 0$,*
i) $\mathrm{P}\big[E_{m,n}^c\big] \leq 8\,\mathrm{e}^{-nH(\delta)}$,
ii) $\mathrm{P}\big[F_{m,n,\varepsilon}^c\big] \leq 10\,\mathrm{e}^{-nH(\delta)}$.

Proof. The Azuma-Hoeffding Theorem [5, 14] says that if $(V_0, \ldots, V_m)$ is a martingale with $V_0 \equiv 0$ and $\mathrm{P}[|V_k - V_{k-1}| \leq a] = 1$ for all $k \in \mathbb{N}_m$ then

$$\mathrm{P}[V_m \geq mb] \leq \exp\left(-\frac{mb^2}{2a^2}\right)$$

for all $b > 0$. For $\xi \in \mathcal{A}_{m,n}$ and $(e_1, e_2) \in \{0, 1\}^2$ fixed let

$$Z_i(\omega) := \begin{cases} 1 & \text{if } \big(X_i(\omega), Y_{\xi(i)}(\omega)\big) = (e_1, e_2), \\ 0 & \text{otherwise.} \end{cases}$$

Then Z_i ($i \in \mathbb{N}_m$ are i.i.d. random variables with expectation $\mathrm{E}[Z_i] = 1/4$. If we set $V_0 := 0$ and

$$V_k := \sum_{i=1}^{k} \big(Z_i - \mathrm{E}[Z_i]\big) \qquad (k \in \mathbb{N}_m),$$

then $(V_0, \ldots, V_m)$ is a martingale with $|V_k - V_{k-1}| \leq 3/4$ for all k. By the Azuma-Hoeffding Theorem, we have

$$\mathrm{P}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i - \frac{1}{4} \geq \epsilon_1(\delta)\right] = \mathrm{P}\left[\frac{1}{m} V_m \geq \epsilon_1(\delta)\right] \leq \exp\left(-\frac{8m\epsilon_1^2(\delta)}{9}\right) \leq \mathrm{e}^{-2nH(\delta)}\,.$$

Applying the same reasoning to the martingale $(-V_0, \ldots, -V_m)$, we find

$$\mathrm{P}\left[\frac{1}{m}\sum_{i=1}^{m} Z_i - \frac{1}{4} \leq -\epsilon_1(\delta)\right] \leq \mathrm{e}^{-2nH(\delta)},$$

so that

$$\mathrm{P}\left[\left|\mathrm{P}_{\widehat{\mathscr{D}}_\xi}\left[(e_1, e_2)\right] - \frac{1}{4}\right| \geq \epsilon_1(\delta)\right] \leq 2\,\mathrm{e}^{-2nH(\delta)}\,.$$

Since this is true for all $(e_1, e_2) \in \{0,1\}^2$, simple union bounds show that

$$\mathrm{P}\left[E_\xi^c\right] \leq 8\,\mathrm{e}^{-2nH(\delta)}$$

and

$$\mathrm{P}\left[E_{m,n}^c\right] = \mathrm{P}\left[\bigcup_{\xi \in \mathcal{A}_{m,n}} E_\xi^c\right] \leq \#\mathcal{A}_{m,n} \times 8\,\mathrm{e}^{-2nH(\delta)} \overset{Lem 2.1}{\leq} 8\,\mathrm{e}^{-nH(\delta)}\,.$$

ii) The proof of the second part is similar: Let (ξ, λ) be such that $\xi \leq \lambda$ and $d(\xi, \lambda) \geq m\varepsilon$. For all i such that $\xi(i) \neq \lambda(i)$ we have $\xi(i) < \lambda(i)$. For $e \in \{-1, 0, 1\}$ fixed let

$$Z_i := \begin{cases} 1 & \text{if } R_i^{\xi,\lambda} = e, \\ 0 & \text{otherwise} \end{cases} \qquad (i \in \{k : \xi(k) \neq \lambda(k)\}.$$

Then we have

$$\mathrm{E}[Z_i] = \begin{cases} \frac{1}{4} & \text{if } e \in \{-1, 1\}, \\ \frac{1}{2} & \text{if } e = 0., \end{cases}$$

so that $|Z_i - \mathrm{E}[Z_i]| \leq 3/4$ in all three cases. Furthermore, the random variables

$$(Z_i - E[Z_i]), \qquad (i \in \{k : \xi(k) \neq \lambda(k)\})$$

are i.i.d. with distribution $\mathscr{J}^{disag}$. This is seen by induction, using the observation that for all index sets

$$I \subset \{k : \xi(k) \neq \lambda(k)\},$$

if $i_{\max} = \max I$ then $X_{i_{\max}}$ and $Y_{\lambda(i_{\max})}$ do not appear in any of the expressions $(X_i, Y_{\xi(i)}, Y_{\lambda(i)})$ $(i \in I \setminus \{i_{\max}\})$, so that independently of the value of $Y_{\xi(i_{\max})}$ (which could have appeared in the above expressions at most once as $Y_{\lambda(i)}$), we have

$$\mathrm{P}\left[Y_{\lambda(i_{\max})} \neq Y_{\xi(i_{\max})}\right] = \frac{1}{2}, \quad \mathrm{P}\left[X_{i_{\max}} \neq Y_{\xi(i_{\max})} = Y_{\lambda(i_{\max})}\right] = \frac{1}{4},$$
$$\text{and} \quad \mathrm{P}\left[X_{i_{\max}} = Y_{\xi(i_{\max})} = Y_{\lambda(i_{\max})}\right] = \frac{1}{4}.$$

We define $V_0 \equiv 0$ and for $k \in \mathbb{N}_{d(\xi,\lambda)}$, $V_k := V_{k-1} + Z_i - \mathrm{E}[Z_i]$, where

$$i = \min\{l \in \mathbb{N}_m : \#\{j \leq l : \xi(j) \neq \lambda(j)\} = k\}.$$

Then $(V_0, \ldots, V_{d(\xi,\lambda)})$ is a martingale, and arguing as above by ways of the Azuma-Hoeffding Theorem, we find

$$\mathrm{P}\left[\left|\mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\xi,\lambda}}[e] - \mathrm{P}_{\mathscr{J}^{disag}}[e]\right| \geq \epsilon_3(\delta,\varepsilon)\right]$$
$$= \mathrm{P}\left[\left|\frac{1}{d(\xi,\eta)} V_{d(\xi,\eta)}\right| \geq \epsilon_3(\delta,\varepsilon)\right] \leq 2\exp\left(-\frac{8d(\xi,\lambda)\epsilon_3^2(\delta,\varepsilon)}{9}\right) \leq 2\,\mathrm{e}^{-3nH(\delta)}.$$

Since this holds for all $e \in \{-1,0,1\}$, we have

$$\mathrm{P}\left[G^c_{\xi,\lambda}\right] \leq 6\,\mathrm{e}^{-3nH(\delta)} \tag{3.2}$$

whenever $d(\xi,\lambda) \geq m\varepsilon$ and $\xi \leq \lambda$.

If it is even the case that $d(\xi,\lambda) > m(1-\varepsilon)$, then we find

$$\mathrm{P}\left[\left|\mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\xi,\lambda}}[e] - \mathrm{P}_{\mathscr{J}^{disag}}[e]\right| \geq \epsilon_4(\delta,\varepsilon)\right] \leq 2\exp\left(-\frac{8d(\xi,\lambda)\epsilon_4^2(\delta,\varepsilon)}{9}\right) \leq 2\,\mathrm{e}^{-3nH(\delta)},$$

and also

$$\left|\mathrm{P}_{\widehat{\mathscr{L}}^{unif}_{\xi,\lambda}}[e] - \mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\xi,\lambda}}[e]\right| \leq 2\varepsilon.$$

Therefore,

$$\mathrm{P}\left[\left|\mathrm{P}_{\widehat{\mathscr{L}}^{unif}_{\xi,\lambda}}[e] - \mathrm{P}_{\mathscr{J}^{unif}}[e]\right| \geq \epsilon_4(\delta,\varepsilon) + 2\varepsilon\right]$$
$$\leq \mathrm{P}\left[\left|\mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\xi,\lambda}}[e] - \mathrm{P}_{\mathscr{J}^{disag}}[e]\right| \geq \epsilon_4(\delta,\varepsilon)\right] \leq 2\,\mathrm{e}^{-3nH(\delta)}.$$

Since this holds for all $e \in \{-1,0,1\}$, we have

$$\mathrm{P}\left[H^c_{\xi,\lambda}\right] \leq 6\,\mathrm{e}^{-3nH(\delta)} \tag{3.3}$$

whenever $d(\xi,\lambda) > m(1-\varepsilon)$ and $\xi \leq \lambda$.

Next, let $\xi \leq \lambda$ be such that $d(\xi,\lambda) \leq m(1-\varepsilon)$, and for $e \in \{-1,1\}$ fixed let

$$Z_i := \begin{cases} 1 & \text{if } R_i^{\xi,\lambda} = e, \\ 0 & \text{otherwise} \end{cases} \qquad (i \in \{k : \xi(k) = \lambda(k)\}).$$

Then $\mathrm{E}[Z_i] = 1/2$ so that $|Z_i - \mathrm{E}[Z_i]| = 1/2$, and

$$(Z_i - E[Z_i]), \qquad (i \in \{k : \xi(k) = \lambda(k)\})$$

are i.i.d. random variables with distribution $\mathscr{J}^{agree}$. Let $V_0 := 0$, and for $k \in \mathbb{N}_{m-d(\xi,\lambda)}$, $V_k := V_{k-1} + Z_i - \mathrm{E}[Z_i]$, where

$$i = \min\{l \in \mathbb{N} : \#\{j \leq l : \xi(j) = \lambda(j)\} = k\}.$$

Then $(V_0, \dots, V_{m-d(\xi,\lambda)})$ is a martingale and the large deviations argument from above shows that

$$\begin{aligned}
\mathrm{P}&\left[\left|\mathrm{P}_{\widehat{\mathscr{L}}^{agree}_{\xi,\lambda}}[e] - \mathrm{P}_{\mathscr{J}^{agree}}[e]\right| \geq \epsilon_2(\delta,\varepsilon)\right] \\
&= \mathrm{P}\left[\left|\frac{1}{m-d(\xi,\lambda)} V_{m-d(\xi,\lambda)}\right| \geq \epsilon_2(\delta,\varepsilon)\right] \leq 2\,\mathrm{e}^{-2(m-d(\xi,\lambda))\epsilon_2^2} \leq 2\,\mathrm{e}^{-3nH(\delta)}.
\end{aligned}$$

Since this holds for both $e \in \{1,-1\}$, we find

$$\mathrm{P}\left[F^c_{\xi,\lambda}\right] \leq 4\,\mathrm{e}^{-3nH(\delta)} \tag{3.4}$$

whenever $d(\xi,\lambda) \leq m(1-\varepsilon)$ and $\xi \leq \lambda$.

Finally, the combination of equations (3.2), (3.3), (3.4) and Lemma 2.1 shows that

$$\begin{aligned}
\mathrm{P}\left[F^c_{m,n}\right] &\leq \sum_{\{(\xi,\lambda):\xi\leq\lambda, m\varepsilon\leq d(\xi,\lambda)\leq m(1-\varepsilon)\}} \left(\mathrm{P}\left[F^c_{\xi,\lambda}\right] + \mathrm{P}\left[G^c_{\xi,\lambda}\right]\right) + \sum_{\{(\xi,\lambda):\xi\leq\lambda, m(1-\varepsilon)<d(\xi,\lambda)\}} \mathrm{P}\left[H^c_{\xi,\lambda}\right] \\
&\leq (\mathcal{A}_{m,n})^2 \times \left(10\,\mathrm{e}^{-3nH(\delta)}\right) \\
&\leq 10\,\mathrm{e}^{-nH(\delta)}.
\end{aligned}$$

□

4. An Ergodic Map. Let us now introduce an ergodic map as follows: Let $T \sim \mathscr{U}(\mathbb{N}_m)$ be a uniform random variable on $\mathbb{N}_m$. By Kolmogorov's theorem we may assume without loss of generality that $(\Omega, \mathscr{F}, \mathrm{P})$ is extended so that T is defined on Ω and independent of the X_i and Y_j. Let us define new random sequences $\tilde{X} = \tilde{X}_1 \dots \tilde{X}_m$ and $\tilde{Y} = \tilde{Y}_1 \dots \tilde{Y}_n$ by setting $\tilde{Y} := Y$, $\tilde{X}_i := X_i$ for all $i \neq T$ and

$$\tilde{X}_T := X_T + 1 \mod 2.$$

In other words, $(\tilde{X}, \tilde{Y})$ is obtained from (X,Y) by flipping one random bit of X and keeping all other entries of X and Y unchanged. The map $(X,Y) \mapsto (\tilde{X}, \tilde{Y})$ is measure-preserving, since $\tilde{X}$ and $\tilde{Y}$ again consist of i.i.d. standard Bernoulli variables. Therefore,

$$\mathrm{E}[\Delta] = 0 \tag{4.1}$$

where $\Delta := S^*(\tilde{X}, \tilde{Y}) - S^*(X,Y)$. The construction in the proof of Lemma 2.3 shows that there exists a $\sigma(X,Y)$-measurable map

$$\begin{aligned}
(\Xi, \Lambda) :\ \Omega &\to \mathcal{A}_{m,n} \times \mathcal{A}_{m,n}, \\
\omega &\mapsto \left(\Xi_\omega, \Lambda_\omega\right)
\end{aligned}$$

such that for all $\omega \in \Omega$, $\Xi_\omega \leq \Lambda_\omega$ and

$$\{i : \Xi_\omega(i) \neq \Lambda_\omega(i)\} = \{i : U_i(\omega) = 1\}.$$

Furthermore, X and Y define the $\sigma(X,Y)$-measurable events $E_{m,n}$, $F_{m,n}$ and the $\sigma(X,Y)$-measurable random variable U introduced in Section 3. The following two lemmas show how these objects affect Δ and will be the key tools in the proof of the main theorem of this paper.

LEMMA 4.1. *For all* $\delta < 5/6$ *and* $\varepsilon > 0$,

$$\mathrm{E}\left[\Delta \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \geq \left[\frac{1}{2} - 3\max\left(\epsilon_4(\delta,\varepsilon)+2\varepsilon, \epsilon_3(\delta,\varepsilon)\right)\right] \times \varepsilon$$
$$+ \min\left[\frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon)+\varepsilon\right), -2\epsilon_2(\delta,\varepsilon)\right] \times (1-\varepsilon).$$

Proof. A key observation is that $Y_{\Xi(T)} \neq Y_{\Lambda(T)}$ implies $\Delta = 1$: Without loss of generality we may assume that $\tilde{X}_T = Y_{\Xi(T)}$, so that $X_T \neq Y_{\Xi(T)}$ and $S^*(X,Y) = \sum_{i\neq T} s(X_i, Y_{\Xi(i)})$. But then we have

$$S^*(X,Y) + 1 \geq S^*(\tilde{X},\tilde{Y}) \geq S(\tilde{X},\tilde{Y};\Xi)$$
$$= \sum_{i\neq T} s(X_i, Y_{\Xi(i)}) + s(\tilde{X}_T, Y_{\Xi(T)}) = S^*(X,Y) + 1,$$

so that $\Delta = 1$ indeed. Likewise, $X_T \neq Y_{\Xi(T)} = Y_{\Lambda(T)}$ implies $\Delta = 1$. Using these facts and the trivial lower bound $\Delta \geq -1$, we have

$$\begin{aligned}
&\mathrm{E}\left[\Delta \,\middle\|\, U \geq m(1-\varepsilon), F_{m,n}\right] \\
&\geq \mathrm{E}\left[1\times \mathrm{P}_{\widehat{\mathscr{L}}^{unif}_{\Xi,\Lambda}}[0] + 1\times \mathrm{P}_{\widehat{\mathscr{L}}^{unif}_{\Xi,\Lambda}}[1] - 1\times \mathrm{P}_{\widehat{\mathscr{L}}^{unif}_{\Xi,\Lambda}}[-1] \,\middle\|\, U \geq m(1-\varepsilon), F_{m,n}\right] \\
&\geq 1\times\left(\mathrm{P}_{\mathscr{J}^{unif}}[0] - \epsilon_4(\delta,\varepsilon) - 2\varepsilon\right) + 1\times\left(\mathrm{P}_{\mathscr{J}^{unif}}[1] - \epsilon_4(\delta,\varepsilon) - 2\varepsilon\right) \\
&\qquad - 1\times\left(\mathrm{P}_{\mathscr{J}^{unif}}[-1] + \epsilon_4(\delta,\varepsilon) + 2\varepsilon\right) \\
&= \frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon\right),
\end{aligned} \tag{4.2}$$

$$\begin{aligned}
&\mathrm{E}\left[\Delta \,\middle\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, F_{m,n}, \Xi(T) \neq \Lambda(T)\right] \\
&\geq \mathrm{E}\Big[1\times \mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\Xi,\Lambda}}[0] + 1\times \mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\Xi,\Lambda}}[1] - 1\times \mathrm{P}_{\widehat{\mathscr{L}}^{disag}_{\Xi,\Lambda}}[-1] \,\Big\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, \\
&\qquad\qquad F_{m,n}, \Xi(T) \neq \Lambda(T)\Big] \\
&\geq 1\times\left(\mathrm{P}_{\mathscr{J}^{disag}}[0] - \epsilon_3(\delta,\varepsilon)\right) + 1\times\left(\mathrm{P}_{\mathscr{J}^{disag}}[1] - \epsilon_3(\delta,\varepsilon)\right) \\
&\qquad - 1\times\left(\mathrm{P}_{\mathscr{J}^{disag}}[-1] + \epsilon_3(\delta,\varepsilon)\right) \\
&= \frac{1}{2} - 3\epsilon_3(\delta,\varepsilon),
\end{aligned} \tag{4.3}$$

$$\begin{aligned}
&\mathrm{E}\left[\Delta \,\middle\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, F_{m,n}, \Xi(T) = \Lambda(T)\right] \\
&\geq \mathrm{E}\Big[1\times \mathrm{P}_{\widehat{\mathscr{L}}^{agree}_{\Xi,\Lambda}}[1] - 1\times \mathrm{P}_{\widehat{\mathscr{L}}^{agree}_{\Xi,\Lambda}}[-1] \,\Big\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, F_{m,n}, \Xi(T) = \Lambda(T)\Big] \\
&\geq 1\times\left(\mathrm{P}_{\mathscr{J}^{agree}}[1] - \epsilon_2(\delta,\varepsilon)\right) - 1\times\left(\mathrm{P}_{\mathscr{J}^{agree}}[-1] + \epsilon_2(\delta,\varepsilon)\right) \\
&= -2\epsilon_2(\delta,\varepsilon),
\end{aligned} \tag{4.4}$$

Putting the pieces together, we find

$$
\begin{aligned}
&\mathrm{E}\left[\Delta \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \\
&\quad \geq \mathrm{E}\left[\Delta \,\middle\|\, U \geq m(1-\varepsilon), F_{m,n}\right] \times \Big(\mathrm{P}\left[U \geq m(1-\varepsilon), \Xi(T) \neq \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \\
&\qquad\qquad\qquad\qquad\qquad + \mathrm{P}\left[U \geq m(1-\varepsilon), \Xi(T) = \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right]\Big) \\
&\quad + \mathrm{E}\left[\Delta \,\middle\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, F_{m,n}, \Xi(T) \neq \Lambda(T)\right] \\
&\qquad\qquad\qquad\qquad\qquad \times \mathrm{P}\left[m(1-\varepsilon) \geq U \geq m\varepsilon, \Xi(T) \neq \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \\
&\quad + \mathrm{E}\left[\Delta \,\middle\|\, m(1-\varepsilon) \geq U \geq m\varepsilon, F_{m,n}, \Xi(T) = \Lambda(T)\right] \\
&\qquad\qquad\qquad\qquad\qquad \times \mathrm{P}\left[m(1-\varepsilon) \geq U \geq m\varepsilon, \Xi(T) = \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right]
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(4.2),(4.3),(4.4)}{\geq} \left[\frac{1}{2} - 3\max\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon, \epsilon_3(\delta,\varepsilon)\right)\right] \times \mathrm{P}\left[\Xi(T) \neq \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \\
&\qquad + \min\left[\frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon\right), -2\epsilon_2(\delta,\varepsilon)\right] \times \mathrm{P}\left[\Xi(T) = \Lambda(T) \,\middle\|\, U \geq m\varepsilon, F_{m,n}\right] \\
&\quad \geq \left[\frac{1}{2} - 3\max\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon, \epsilon_3(\delta,\varepsilon)\right)\right] \times \varepsilon + \min\left[\frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon\right), -2\epsilon_2(\delta,\varepsilon)\right] \times (1-\varepsilon).
\end{aligned}
$$

□

LEMMA 4.2. *For any $\sigma(X,Y)$-measurable event B and all $\delta < 5/6$, we have*

$$\int_B \Delta \, d\,\mathrm{P} \geq -4\epsilon_1(\delta) - 8\,\mathrm{e}^{-nH(\delta)}.$$

Proof.

$$\int_B \Delta \, d\,\mathrm{P} \geq \int_{B\cap E_{m,n}} \Delta \, d\,\mathrm{P} - \mathrm{P}[E^c_{m,n}] \overset{Lem 3.2}{\geq} \int_{B\cap E_{m,n}} \Delta \, d\,\mathrm{P} - 8\,\mathrm{e}^{-nH(\delta)}. \tag{4.5}$$

Clearly, for all $\omega \in \Omega$,

$$\Delta(\omega) \geq S\big(\tilde{X}(\omega), \tilde{Y}(\omega); \Xi_\omega\big) - S\big(X(\omega), Y(\omega); \Xi_\omega\big).$$

Therefore,

$$
\begin{aligned}
\int_{B\cap E_{m,n}} \Delta \, d\,\mathrm{P} &\geq \int_{B\cap E_{m,n}} \left[S\left(\tilde{X}, \tilde{Y}; \Xi\right) - S\left(X, Y; \Xi\right)\right] d\,\mathrm{P} \\
&= \int_{B\cap E_{m,n}} \left[s\left(\tilde{X}_T, Y_{\Xi(T)}\right) - s\left(X_T, Y_{\Xi(T)}\right)\right] d\,\mathrm{P}\left[(X(\omega), Y(\omega), T(\omega))\right] \\
&= \int_{B\cap E_{m,n}} \mathrm{E}\left[s\left(\tilde{X}_T, Y_{\Xi(T)}\right) - s\left(X_T, Y_{\Xi(T)}\right) \,\middle\|\, (X,Y)\right] d\,\mathrm{P}\left[(X(\omega), Y(\omega))\right] \\
&\overset{(3.1)}{\geq} \int_{B\cap E_{m,n}} -1 \times 4\epsilon_1(\delta)\, d\,\mathrm{P}\left[(X(\omega), Y(\omega))\right] \geq -4\epsilon_1(\delta).
\end{aligned}
$$

Together with (4.5) this implies the claim. □

5. Proof of The Main Theorem. After introducing the tools of Sections 3 and 4, the stage is set for a proof of Theorem 1.1.

Proof. Since $\{\omega : U < m\varepsilon\} \cup F^c_{m,n}$ is $\sigma(X,Y)$-measurable, we have

$$0 = \mathrm{E}[\Delta] = \mathrm{E}\left[\Delta \middle\| U \geq m\varepsilon, F_{m,n}\right] \times \mathrm{P}\left[U \geq m\varepsilon, F_{m,n}\right] + \int_{\{U<m\varepsilon\}\cup F^c_{m,n}} \Delta d\,\mathrm{P} \overset{Lem 4.1, Lem 4.2}{\geq}$$

$$\geq \left(\left[\frac{1}{2} - 3\max\left(\epsilon_4(\delta,\varepsilon) + 2\varepsilon, \epsilon_3(\delta,\varepsilon)\right)\right] \times \varepsilon + \min\left[\frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon)+2\varepsilon\right), -2\epsilon_2(\delta,\varepsilon)\right] \times (1-\varepsilon)\right)$$
$$\times \left(\mathrm{P}\left[U \geq m\varepsilon\right] - \mathrm{P}\left[F^c_{m,n}\right]\right)$$
$$-\left(4\epsilon_1(\delta) + 8\,\mathrm{e}^{-nH(\delta)}\right).$$

Therefore,

$$\mathrm{P}\left[U \geq m\varepsilon\right]$$
$$\leq \frac{4\epsilon_1(\delta) + 8\,\mathrm{e}^{-nH(\delta)}}{\left[\frac{1}{2} - 3\max\left(\epsilon_4(\delta,\varepsilon)+2\varepsilon, \epsilon_3(\delta,\varepsilon)\right)\right] \times \varepsilon + \min\left[\frac{1}{2} - 3\left(\epsilon_4(\delta,\varepsilon)+2\varepsilon\right), -2\epsilon_2(\delta,\varepsilon)\right] \times (1-\varepsilon)}$$
$$+ 10\,\mathrm{e}^{-nH(\delta)} = \frac{\mathcal{O}\left(\delta^{\frac{1}{2}}\right)}{\mathcal{O}(\varepsilon) + \mathcal{O}\left(\delta^{\frac{1}{2}}\right)} + \mathcal{O}\left(\mathrm{e}^{-n\delta}\right).$$

∎

REFERENCES

[1] D. Aldous and P. Diaconis. Longest increasing subsequences: from patience sorting to the Baik-Deift-Johansson theorem. *Bull. Amer. Math. Soc. (N.S.)*, 36(4):413–432, 1999.

[2] K.S. Alexander. The rate of convergence of the mean length of the longest common subsequence. *Ann. Appl. Probab.*, 4(4):1074–1082, 1994.

[3] S. Amsalu, R.A. Hauser, and H. Matzinger. A Monte Carlo approach to the fluctuation problem in optimal alignments of random strings. arXiv:1211.5489, 2012.

[4] R. Arratia and M.S. Waterman. A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Probab.*, 4(1):200–225, 1994.

[5] K. Azuma. Weighted sums of certain dependent random variables. *Tohuku Math. J.*, 19:357–367, 1967.

[6] R.A. Baeza-Yates, R. Gavaldà, G. Navarro, and R. Scheihing. Bounding the expected length of longest common subsequences and forests. *Theory Comput. Syst.*, 32(4):435–452, 1999.

[7] J. Baik, P. Deift, and K. Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 12(4):1119–1178, 1999.

[8] F. Bonetto and H. Matzinger. Fluctuations of the longest common subsequence in the case of 2- and 3-letter alphabets. *Latin American Journal of Probability and Mathematics*, 2:195–216, 2007.

[9] V. Chvatal and D. Sankoff. Longest common subsequences of two random sequences. *J. Appl. Probability*, 12:306–315, 1975.

[10] V. Dancik and M. Paterson. Upper bounds for the expected length of a longest common subsequence of two binary sequences. *Random Structures Algorithms*, 6(4):449–458, 1995.

[11] J.G. Deken. Some limit results for longest common subsequences. *Discrete Math.*, 26(1):17–31, 1979.

[12] R.A. Hauser, S. Martinez, and H. Matzinger. Large deviation based upper bounds for the LCS problem. *Advances in Applied Probability*, 38:827852, 2006.

[13] E. Hirmo, J. Lember, and H. Matzinger. Detecting the homology of DNA-sequences based on the variety of optimal alignments: a case study. arXiv:1210.3771 [stat.AP], 2012.

[14] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statis. Assoc.*, 58:13–30, 1963.

[15] C. Houdre, J. Lember, and H. Matzinger. On the longest common increasing binary subsequence. *C.R. Acad. Sci. Paris, Ser. I*, 343:589594, 2006.

[16] J. Lember and H. Matzinger. Standard deviation of the longest common subsequence. *Annals of Probability*, 37:11921235, 2009.

[17] J. Lember, H. Matzinger, and C. Dürringer. Deviation from the mean in sequence comparison with a periodic sequence. *Alea*, 3:129, 2007.

[18] M. Paterson and V. Dancik. Longest common subsequences. In *Mathematical foundations of computer science 1994 (Kosice, 1994)*, volume 841 of *Lecture Notes in Comput. Sci.*, pages 127–142. Springer, Berlin, 1994.

[19] H. Robbins. A remark on Stirling's formula. *Amer. Math. Monthly*, 62:26–29, 1955.

[20] M.J. Steele. An Efron-Stein inequality for nonsymmetric statistics. *Ann. Statist.*, 14:753–758, 1986.

[21] M.S. Waterman. General methods of sequence comparison. *Bull. Math. Biol.*, 46(4):473–500, 1984.

[22] M.S. Waterman. Estimating statistical significance of sequence alignments. *Phil. Trans. R. Soc. Lond. B*, 344:383–390, 1994.